# The mosaic genome structure and phylogeny of Shiga toxin-producing Escherichia coli O104:H4 is driven by short-term adaptation

Zhou, K; Ferdous, M; de Boer, R F; Kooistra-Smid, A M D; Grundmann, H; Friedrich, A W; Rossen, J W A

*Published in:*
Clinical Microbiology and Infection

*DOI:*
10.1016/j.cmi.2014.12.009



*Publication date:*
2014

# The mosaic genome structure and phylogeny of Shiga toxin-producing *Escherichia coli* O104:H4 is driven by short-term adaptation

**K. Zhou[1], M. Ferdous[1], R. F. de Boer[2], A. M. D. Kooistra-Smid[1,2], H. Grundmann[1], A. W. Friedrich[1] and J. W. A. Rossen[1]**

1) *University of Groningen, University Medical Center Groningen and* 2) *Certe Laboratory for Infectious Diseases, Groningen, The Netherlands*

## Abstract

Shiga toxin-producing *Escherichia coli* (STEC) O104:H4 emerged as an important pathogen when it caused a large outbreak in Germany in 2011. Little is known about the evolutionary history and genomic diversity of the bacterium. The current communication describes a comprehensive analysis of STEC O104:H4 genomes from the 2011 outbreak and other non-outbreak-related isolates. Outbreak-related isolates formed a tight cluster that shared a monophyletic relation with two non-outbreak clusters, suggesting that all three clusters originated from a common ancestor. Eight single nucleotide polymorphisms, seven of which were non-synonymous, distinguished outbreak from non-outbreak isolates. Lineage-specific markers indicated that recent partitions were driven by selective pressures associated with niche adaptation. Based on the results, an evolutionary model for STEC O104:H4 is proposed. Our analysis provides the evolutionary context at population level and describes the emergence of clones with novel properties, which is necessary for developing comprehensive approaches to early warning and control.



**Keywords:** Antibiotic resistance, comparative genomics, *Escherichia coli* O104:H4, genomic islands, genomic structural variation, next-generation sequencing, prophages, shiga toxin-producing *Escherichia coli*, single nucleotide polymorphism, Stx2-encoding prophage
**Original Submission:** 15 October 2014; **Revised Submission:** 17 December 2014; **Accepted:** 17 December 2014
Editor: F. Allerberger
**Article published online:** 27 December 2014

**Corresponding author:** K. Zhou, Department of Medical Microbiology, University Medical Center Groningen, hpc EB 80, Hanzeplein 1, 9713 GZ Groningen, The Netherlands
**Corresponding author:** A.W. Friedrich, Department of Medical Microbiology, University Medical Centre Groningen, hpc EB 80, Hanzeplein 1, 9713 GZ Groningen, The Netherlands
**E-mails:** K.Zhou@umcg.nl **(K. Zhou),** Alex.Friedrich@umcg.nl **(A.W. Friedrich)**

## Introduction

From early May to July 2011, nearly 4000 clinical infections were ascertained by health authorities in Germany accounting for the largest Shiga toxin-producing *Escherichia coli* (STEC) outbreak on record. Over 900 patients developed haemolytic uraemic syndrome (HUS), of which 54 died [1]. Two features set this outbreak apart from previous ones caused by STEC O157:H7, including the high incidence of HUS (>20%) and a rare serotype O104:H4 [1]. Isolates associated with the outbreak had an unusual combination of virulence factors not only attributed to STEC but also to enteroaggregative *E. coli* (EAEC) harbouring the Stx2-encoding prophage and pAA-like plasmid [2], which may have contributed to the high rate of HUS [3]. Moreover, outbreak isolates contained an extended spectrum β-lactamase (ESBL) gene, which is rare in STEC [4].

Using the power of next-generation sequencing technology, the first available draft sequence of an outbreak strain (TY-2482) isolated from a 16-year-old girl became available while the outbreak was still ongoing. It revealed a high degree of genome plasticity with numerous mobile genetic elements (MGEs) and three plasmids [2]. Further analysis showed that outbreak strains shared the same sequence type (ST) known as ST678 with a historical STEC O104:H4 strain 01-09591, which was isolated from a child presenting with HUS in Germany in 2001. Genomic comparisons revealed a genetic relationship of



http://dx.doi.org/10.1016/j.cmi.2014.12.009

99.8% nucleotide similarity with an AggR-positive EAEC O104:H4 strain 55989 isolated in Central Africa in the late 1990s [2]. This was further supported by a study that included additional EAEC O104:H4 strains in the phylogenetic analysis. It was therefore suggested that EAEC O104:H4 strain 55989 represented a clade at the root of the emerging clone of STEC O104:H4 that rapidly expanded in 2011 [5]. The limited number of single nucleotide polymorphisms (SNPs) among all sequenced outbreak isolates suggested their clonality [6,7]. However, it remains to be elucidated how this clone evolved and attained its repertoire of virulence factors. In this study, we attempt to shed light on the evolution of STEC O104:H4 by describing the genome structure and population structure of outbreak isolates and available non-outbreak isolates obtained from sporadic infections reported before and after the outbreak.

## Materials and methods

### Strains analysed in this study

In all, 23 *E. coli* isolates have been used in this study (Table 1). Seven isolates were sequenced as part of our previous study (Ferdous et al., unpublished). Briefly, DNA libraries were prepared using the Nextera XT v2 kit (Illumina, San Diego, CA, USA) according to the manufacturer's instructions and then run on a Miseq (Illumina) for generating paired-end 250-bp reads. *De novo* assembly was performed using CLC Genomics Workbench v6.0.5 (CLC bio A/S, Aarhus, Denmark) after quality trimming (Qs $\geq$28) with optimal word sizes based on the maximum N50 value. Four of the seven isolates were obtained from a HUS patient (338) and her travel partner (381-1, 381-3 and 381-4), and the other three (7N, 8G and 9Z) were isolated during the 2011 outbreak in Germany. Apart from isolate 381-3, which is non-O104:H4, *stx* negative and ST10, the other six isolates belong to STEC O104:H4/ST678. Therefore 381-3 was only used for plasmid comparison in this study. In addition, 17 genomes were obtained from publically available databases: genome sequences of 55989, 2009EL-2050, 2009EL-2071, 2011C-3493 and E112/10 were downloaded from the NCBI database, and the others were downloaded from http://www.broadinstitute.org/annotation/genome/Ecoli_O104_H4/Downloads.html. Detailed information of all the isolates analysed in this study is listed in Table 1. The GenBank accession numbers of all isolates analysed in this study were the following: NC_011748 (55989), NC_018650 (2009EL-2050), NC_018661 (2009EL-2071), NC_018658 (2011C-3493), NZ_AHAV00000000 (E112/10), AFVR00000000 (TY-2482), AFUX01000000 (Ec11-4404), AFVA01000000 (Ec11-4632.1), AIPQ01000000 (Ec12-0465), AIPR01000000 (Ec12-0466), AGWF01000000 (Ec11-9450), AGWH01000000 (Ec11-9941), AGWG01000000 (Ec11-9990), AFRL01000000 (04-8351), AFRK01000000 (09-7901), AFPS00000000 (HUSEC041), JRJF00000000 (338), JRKD00000000 (381-1), JRLM00000000 (381-3), JRLD00000000 (381-4), JRKE00000000 (7N), JRLN00000000 (8G), JRKF00000000 (9Z).

### Annotation

To annotate the genomes, contigs were first oriented and ordered using ABACAS [8] against the reference TY-2482 chromosome and plasmids with the following settings: using sensitive mapping in Numer, a minimum per cent identity of 40, a minimum per cent contig coverage of 20, minimum contig converage difference set to 0, and reference sequence is circular. The start coordinate of all genomes has been reset according to the first nucleotide of TY-2482 (GATGTTGCTCCCCCAAG). Contigs were concatenated following this order as a pseudomolecule with appending the unmapped contigs at the end. Each ordered genome was manually curated after performing automatic annotation on the RAST server [9].

### Mapping and SNP analysis

Reads were mapped to the chromosome of TY-2482 by CLC Genomics Workbench v6.05 with default settings. To acquire reliable SNPs, the regions of MGEs (prophages and genomic islands) and repeats were masked during mapping. Candidate SNPs were called by the algorithm Quality-based variant detection of CLC Genomics Workbench. SNPs were filtered out if one of the following occurred: (i) their quality score was <30; (ii) the neighbourhood quality was <30; (iii) the minimum variant frequency was <35%; (iv) the minimum coverage was <10; (v) they were only detected on a single strand. SNPs called from assembly genomes were identified by Mauve [10].

### Genome analysis: genomic islands, prophages and plasmids

Fragments >5 kb that were absent in at least one genome were detected by BLAST and were defined as genomic islands (GEIs) in this study. The prophages were predicted on the web server PHAST [11] followed by manual curations. Only 'intact' prophages detected by PHAST were included in the further analysis, and those were grouped according to the sequence similarity aligned by Mauve. The plasmid analysis was mainly dependent on BLASTn. The contigs of each sample were blasted against the reference plasmid and plotted by BLAST Ring Image Generator (BRIG) [12]. The reference plasmid was artificially generated by concatenating sequences of a set of plasmids, including pTY1, pTY2, pTY3 [2], p55989 [13], pHUSEC41-1, pHUSEC41-2, pHUSEC41-3, pHUSEC41-4 [14] and p09EL50 [15].

**TABLE 1. Isolates analysed in this study [30–33]**

| Isolate ID[a] | Date of isolate | Patient information | Clinical manifestations | Epidemic information | Country of isolation | ESBL | Virulence group[b] | Reference |
|---|---|---|---|---|---|---|---|---|
| 7N | 2011 | unknown | unknown | German outbreak | Germany | + | Group I | Ferdous *et al.*, unpublished |
| 8G | 2011 | unknown | unknown | German outbreak | Germany | + | Group I | Ferdous *et al.*, unpublished |
| 9Z | 2011 | unknown | HUS | German outbreak | Germany | + | Group I[c] | Ferdous *et al.*, unpublished |
| TY-2482 | 2011 | 16-year-old female | HUS | German outbreak | Germany | + | Group I | [2] |
| 2011c-3493 | 2011 | 51-year-old male | HUS | Germany, travel, German outbreak period | U.S. | + | Group I | [15] |
| Ec11-4404 | 06. 2011 | male | HUS | French outbreak | France | + | Group I | [7] |
| Ec11-4632.1 | 06. 2011 | female | HUS | French outbreak | France | + | Group I | [7] |
| Ec12-0466 | 12. 2011 | child | HUS | North Africa, travel | France | – | Group I | [18] |
| 381-4 | 07. 2013 | 23-year-old female | diarrhoea | Turkey, travel | Netherlands | + | Group I | Ferdous *et al.*, unpublished |
| 381-1 | 07. 2013 | 23-year-old female | diarrhoea | Turkey, travel | Netherlands | – | Group I | Ferdous *et al.*, unpublished |
| 338 | 07. 2013 | 22-year-old female | HUS | Turkey, travel | Netherlands | – | Group I | Ferdous *et al.*, unpublished |
| Ec11-9941 | 9.2011 | child | HUS | unknown | France | – | Group I | [18] |
| E112/10 | 2010 | unknown | unknown | Tunisia, travel | Sweden | – | Group I | [18] |
| Ec11-9990 | 8.2011 | child | HUS | Unknown | France | – | Group I | [18] |
| Ec11-9450 | 10. 2011 | unknown | HUS | Turkey, travel, local outbreak | France | – | Group I[c] | [30] |
| 2009EL-2071 | 2009 | unknown | bloody diarrhoea | unknown | Republic of Georgia | – | Group I | [31] |
| Ec12-0465 | 11. 2011 | child | HUS | unknown | France | – | Group I | [18] |
| 2009EL-2050 | 2009 | unknown | bloody diarrhoea | unknown | Republic of Georgia | – | Group I | [31] |
| 04-8351 | 2004 | 6-year-old male | haemorrhagic colitis | unknown | France | – | Group II | [32] |
| 09-7901 | 2009 | adult male | HUS | unknown | France | – | Group II | [32] |
| HUSEC041 (01-09591) | 2001 | child | HUS | unknown | Germany | – | Group II | [33] |
| 55989 | Late 1990s | HIV patient | diarrhoea | Unknown | Central African Republic | – | Group III | [13] |
| 381-3[d] | 07. 2013 | 23-year-old female | diarrhoea | Turkey, Travel | Netherlands | + | Group IV | Ferdous *et al.*, unpublished |

[a] The isolates listed here were grouped in different colours according to the phylogenetic results shown in Fig. 1. The sequence type and serotype of all isolates is ST678 and O104:H4, with the exception of isolate 381-3, which is ST-10 and O126:H2.

[b] The virulence groups are defined as Group I (positive for *stx2*/*aggA*/*aggR*/*aatA*/*sigA*/*pic*/*iha*), Group II (positive for *stx2*/*agg3A*/*aggR*/*aatA*/*sigA*/*pic*/*iha*), Group III (positive for *agg3A*/*aggR*/*aatA*/*sigA*/*pic*/*iha*) and group IV (positive for *aatA*/*iha*).

[c] Strain 9Z lost a fragment containing *aggR* (please refer to the text for more details), and strain Ec11-9450 lost pTY2 *in vitro* as described previously [18].

[d] This strain was not included in the phylogenetic analysis but only in the plasmid analysis.

### Core-genome phylogenetic analysis

The whole genomes were aligned using Mauve. Fragments (≥500 bp) shared by all genomes were collected and then concatenated. The resulting pseudomolecules were defined as the core genome, which was used for the phylogenetic analysis. SNPs were collected from the core genomes by in-house scripts. A maximum likelihood (ML) phylogeny was estimated by RAxML v7.2.8 [16] with 1000 bootstrap replications under the general time-reversible model with Gamma correction (GTR+G).

## Results

### Core-genome phylogeny of STEC O104:H4

To reveal the evolutionary relationship of STEC O104:H4 analysed in this study, a core-genome phylogenetic analysis based on SNPs was performed. An ML phylogenetic tree was constructed based on 3659 SNPs detected from the alignments of the 4.5 Mbp core genome (Fig. 1). The phylogeny showed that the sequenced German outbreak isolates 7N, 8G and 9Z from our previous study (Ferdous et al., unpublished) shared a monophyletic relationship (outbreak clade; highlighted in red in Fig. 1) with two other German outbreak isolates (TY-2482 and 2011C-3493) and two French outbreak isolates (Ec11-4404 and Ec11-4632.1). Three isolates from 2013 (338, 381-1 and 381-4) clustered in a separated clade (non-outbreak clade A, abbreviated to clade A; highlighted in green in Fig. 1) together with four other non-outbreak isolates (E112/10, Ec11-9941, Ec11-9990 and Ec12-0466). E112/10 was isolated in 2010 from a Swedish patient, and Ec11-9941, Ec11-9990 and Ec12-0466 were isolated after the outbreak in France 2011. Notably, Ec12-0466 formed a separated branch within this clade. Two additional 2011 isolates from France (Ec11-9450 and Ec12-

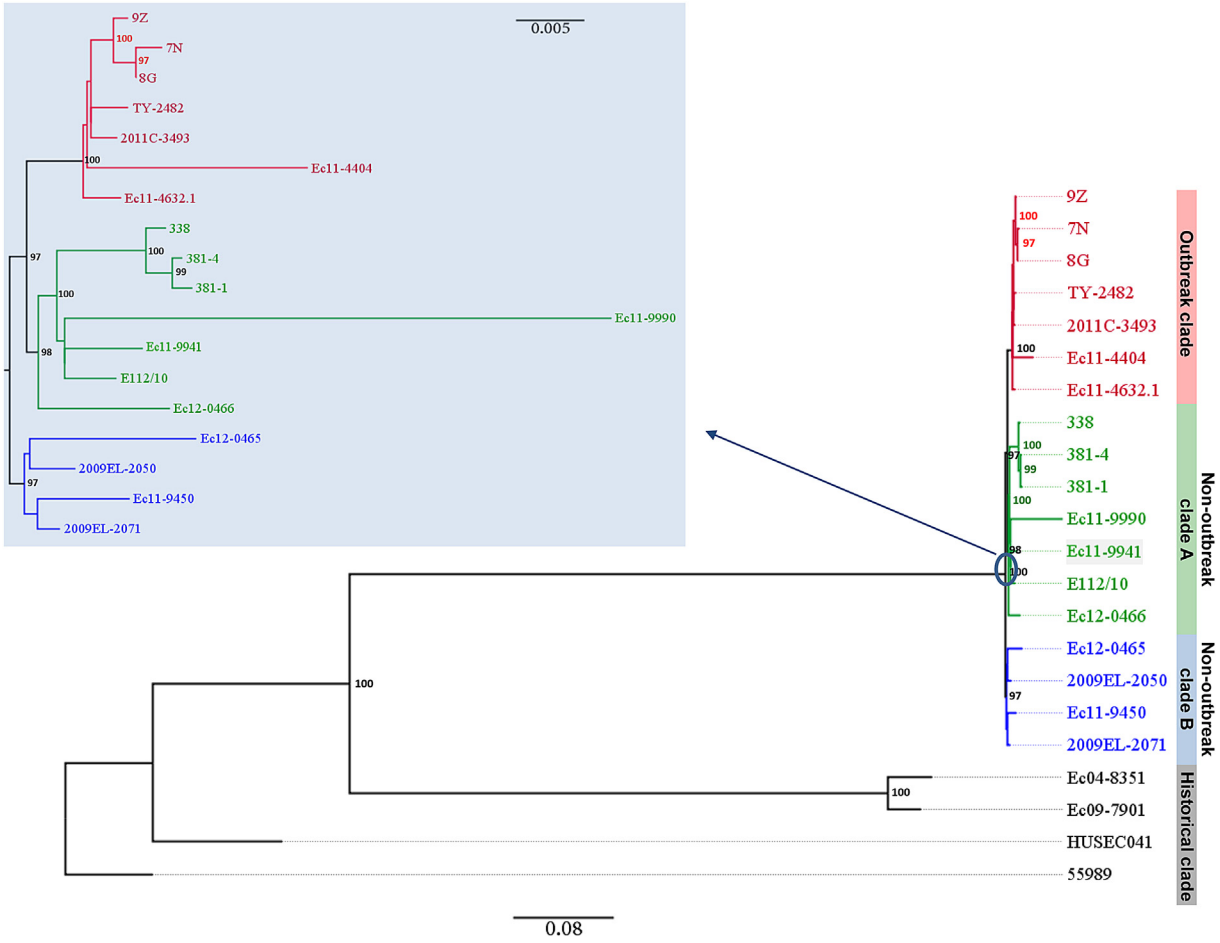


**FIG. 1.** Maximum-likelihood phylogeny of *Escherichia coli* O104:H4. The phylogeny was derived by core-genome analysis using an approximately 4.5-Mbp genome sequence of each sample. The three major clades were respectively referred as to outbreak clade (red), non-outbreak clade A (green), and non-outbreak clade B (blue). The other clades were collectively named 'historical clade' (black). The inset shows the close-up phylogenetic tree of the three major clades. The numbers on the nodes represent the percentage of bootstrap support (>90).

0465) isolated after the outbreak clustered with two 2009 isolates from the Republic of Georgia (2009EL-2050 and 2009EL-2071) forming another distinct clade (non-outbreak clade B, abbreviated to clade B; highlighted in blue in Fig. 1). All three clades are closely related, suggesting that they shared a common ancestor. Notably, clade A and the outbreak clade were more closely related to each other than to clade B (Fig. 1). The three clades share a relatively distant relationship with three isolates (HUSEC041, 04-8351, 09-7901) and the hypothetical progenitor EAEC strain 55989, and the clades formed by the four strains were collectively named as 'historical clade' (shown in black in Fig. 1). Taken together, the phylogeny of STEC O104:H4 indicated that this bacterium has diversified into multiple lineages, at least three of them sharing a close relationship that may represent the dominant population of STEC O104:H4. We note that clade A and B isolates were obtained from different geographic regions (Table 1), indicating the local expansion of certain STEC O104:H4 clones.

### Clade-specific SNPs

We identified eight canonical SNPs in the core genome that are unique to the outbreak clade (Table 2), suggesting that they were acquired by the outbreak clone recently. Mapping these SNPs to the available sequences of 40 additional outbreak isolates (including German and French isolates) reported previously [6,7,17] (http://www.hpa-bioinformatics.org.uk/lgp/genomes) supported their canonical nature. All SNPs located within coding regions and seven of them were non-synonymous.

### Comparison of the accessory genome

*Plasmids.* Plasmids of outbreak strain TY-2482 (pTY1, pTY2 and pTY3), non-outbreak strain 2009EL-2050 (p09EL50) and historical strain HUSEC041 (pHUSEC41-1, pHUSEC41-3 and pHUSEC41-4) were used as reference to investigate the plasmid content of isolates analysed here. Substantial variations in the content of plasmids were observed among strains analysed (Fig. 2). Plasmid pTY1 carrying the ESBL gene $bla_{CTX\text{-}M\text{-}15}$ and a β-lactamase gene $bla_{TEM\text{-}1}$ is present in all isolates of the outbreak clade, but not in any isolates of other clades, indicating that pTY1 may be recently acquired by the outbreak strains resulting in potential adaptive advantages (e.g. antibiotic resistance). The plasmid pTY2 carries an *agg* operon encoding AAF/I fimbriae resulting in the enteroaggregative phenotype of the outbreak strains. The pTY2-like plasmid was not found in any isolates of the historical clade, but in all isolates of outbreak and non-outbreak clades A and B except Ec11-9450 (due to plasmid loss during culturing; [18]). Notably, we observed a spontaneous deletion in the pTY2 plasmid of isolate 9Z resulting in the loss of the *aggR* gene, which encodes a transcriptional activator for the fimbriae expression (see Supporting information, Fig. S1). This gene was detected in the original isolate [3], and may therefore have been lost during propagation *in vitro*. In contrast, another enteroaggregative plasmid p55989 (also known as pAA from EAEC) encoding AAF/III fimbriae instead of AAF/I fimbriae was exclusively found in strains of historical clade, suggesting a recent replacement of pAA by pTY2. Plasmid pTY3 is a small cryptic plasmid only carrying a *repA* gene, which was found in all isolates of outbreak and non-outbreak clades except in Ec11-9990. It was not present in the isolates of historical clade. We used BLAST on the sequence of pTY3 in GenBank to explore the origin of the small cryptic plasmid. Besides plasmids found in *E. coli* O104:H4, highly similar plasmids (identity >90%) were found in other *E. coli* strains and also in some other bacterial species (see Supporting information, Table S1). Therefore, the origin of the small cryptic plasmid could not yet be resolved.

The plasmid pHUSEC41-1 from the historical isolate HUSEC041 carries a *Tn3*-like transposase flanked by the multiple drug-resistance (MDR) genes $bla_{TEM\text{-}1}$, *strA*, *strB* and *sul2*. Besides HUSEC041, pHUSEC41-1-like plasmid was found in historical isolate 04-8351, clade A isolates 381-1, E112/10, Ec11-9941, Ec11-9990, Ec12-0466, clade B isolates Ec11-9960 as well as the non-O104/*stx*-negative isolate 381-3 (Fig. 2). However, none of the outbreak isolates harboured this plasmid, which may be caused by the fact that both plasmid pTY1 and pHUSEC41 share the same incompatibility group (IncI1). Notably, the region containing MDR genes was missing on the

**TABLE 2. Eight SNPs distinguishing outbreak isolates from non-outbreak isolates**

| Reference position[a] | SNP (Ob→Nob)[b] | Location | Annotation | Amino acid change |
|---|---|---|---|---|
| 347122 | C→T | CDS | putative oxidoreductase | Arg130Gln |
| 1323393 | A→G | CDS | PTS system, galactitol-specific IIC component GatC | synonymous |
| 1449640 | T→C | CDS | ferredoxin-type protein NapG (periplasmic nitrate reductase) | His47Arg |
| 1768361 | T→G | CDS | uracil phosphoribosyltransferase protein | Glu184Asp |
| 2602394 | A→G | CDS | putative calcium/sodium:proton antiporter YrbG | Ile108Met |
| 3033847 | T→G | CDS | selenocysteine-specific translation elongation factor | Asn169His |
| 3429136 | T→C | CDS | rhamnulokinase | Glu424Gly |
| 4527390 | C→T | CDS | DNA-binding ATP-dependent protease La Type I | Thr319Ile |

[a]TY-2482 was used as reference here, of which the start coordinate was reset as described in the text.
[b]Ob and Nob represent outbreak and non-outbreak, respectively.

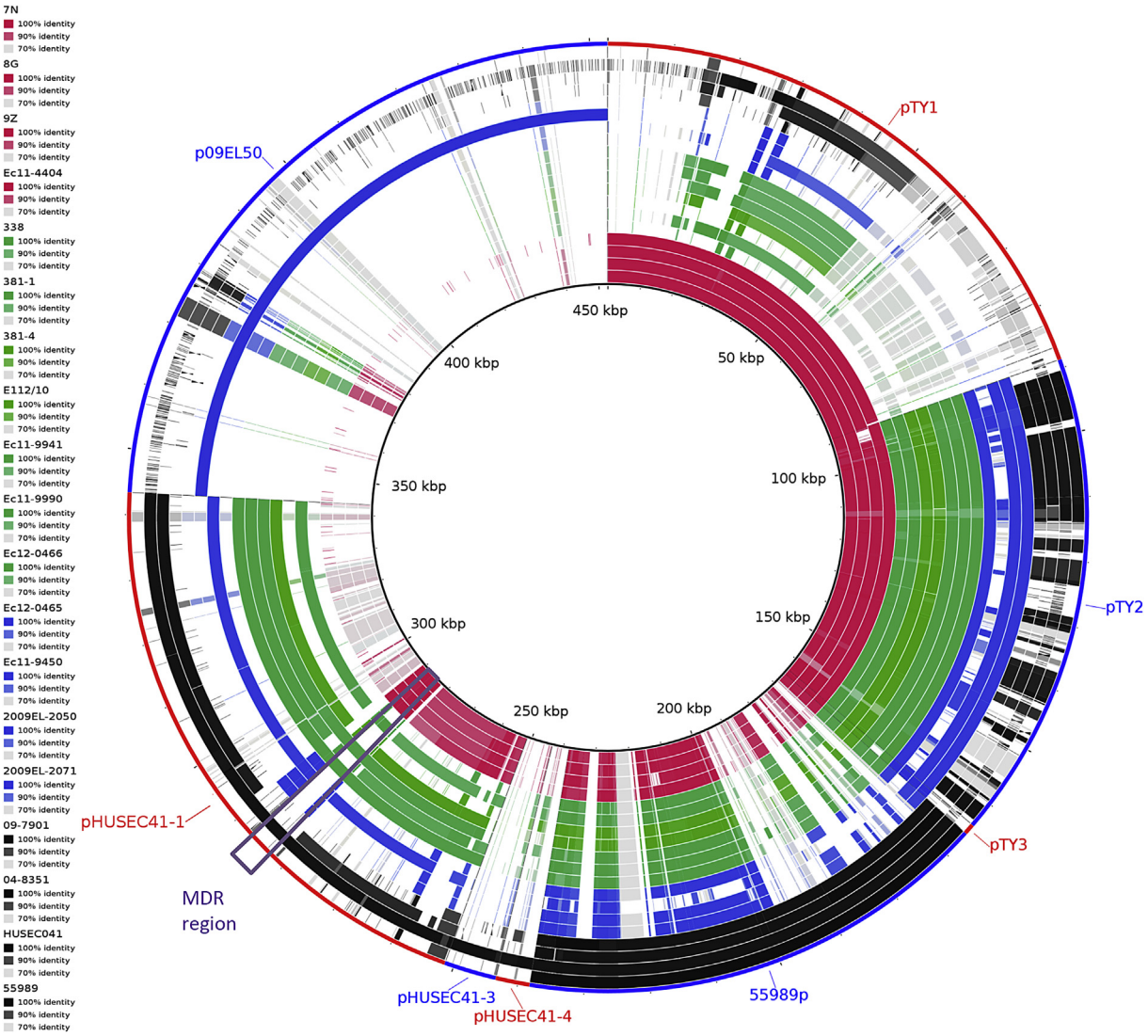


**FIG. 2.** Comparison of the plasmid content in *Escherichia coli* O104:H4 strains. Each ring corresponds to the BLASTn result of one genome relative to the artificial plasmid reference. The reference was composed of numerous plasmids shown by the first outer ring with labels in alternate colours. From outer to inner, the rings were ordered as the sequence shown in the legends (left). Strains were grouped in different colours according to the phylogenetic results shown in Fig. 1. The gradients (dark, pale and white) of each colour represent the sequence similarity (from 100% to 0%) between samples and reference. The multiple drug-resistance region in pHUSEC41-1 is marked by a purple frame.

pHUSEC41-1-like plasmid in 04-8351, Ec11-9450, Ec11-9990 and E112/10 (Fig. 2). However, this region was replaced by another carrying the ESBL gene $bla_{CTX\text{-}M\text{-}15}$ on the pHUSEC41-1-like plasmid of 381-1. To our knowledge, this is the first report of an ESBL-producing non-outbreak isolate. It is noteworthy that an almost identical pHUSEC-41-1-like plasmid as that observed in 381-1 was found in the non-O104/*stx*-negative isolate 381-3, both of which were recovered from the same patient (see Supporting information, Fig. S2). This may result from a possible transconjugation event between 381-1 and 381-3 or between a common donor and both isolates, as the plasmid harboured an intact transconjugation operon (*trb*, *tra* and *pil*). This finding may explain why only 381-1 but not 338 and 381-4 were ESBL positive, although the three isolates were clonal. No significant hits of pHUSEC41-3, pHUSEC41-4 and p09EL50 were found in any of the isolates studied here except their origins.

*Prophages.* Frequent gain or loss of prophages occurred across the investigated population. To further analyse the diversity of prophages among STEC O104:H4 isolates, we used the seven prophages identified from TY-2482 (named as Phage-1 to Phage-

VII according to their positions on the chromosome) [18] as reference to group the others according to sequence identity.

Our analysis found several lineage-specific prophages. Phage-IV was the most diverse prophage found in this study, which was identified in all isolates except for two isolates of the historical clade Ec09-7901 and 55989 (Fig. 3). All Phage-IV shared the same integration site within the *yecE* gene (see Supporting information, Fig. S3). Phylogenetic analysis using ML trees revealed a striking topological homology with the core-genome ML tree indicative of co-evolution. Hence the phage-IV of outbreak isolates clustered tightly in a single clade distant from the clade formed by other isolates, with the exception of Ec12-0466, which appeared to be more closely related to that of the outbreak isolates (Fig. 4a), consistent with its outlier position in the core-genome ML tree. This indicates that a replacement of Phage-IV occurred in the outbreak clone recently, although it remains unknown whether this prophage is functional or not.

Phage-VII carries the *stx2* gene, and so is known as the Stx2-encoding prophage. Excepting the progenitor strain 55989, all other strains harboured this prophage, which chromosomally located within *wrbA*. Remarkably, phylogenetic analysis revealed that Stx2-encoding prophages detected from clade B isolates clustered in a single clade separated from the one formed by all other isolates (Fig. 4b). This suggests that a single replacement of the Stx2-encoding prophage occurred in the ancestor of clade B. Further sequence analysis showed that one of the significant differences between the two clusters of Stx2-encoding prophages was found within the lysis region, where a *rha* gene (encoding the Rha family phage regulatory protein) and an unnamed gene (encoding the lytic protein) were

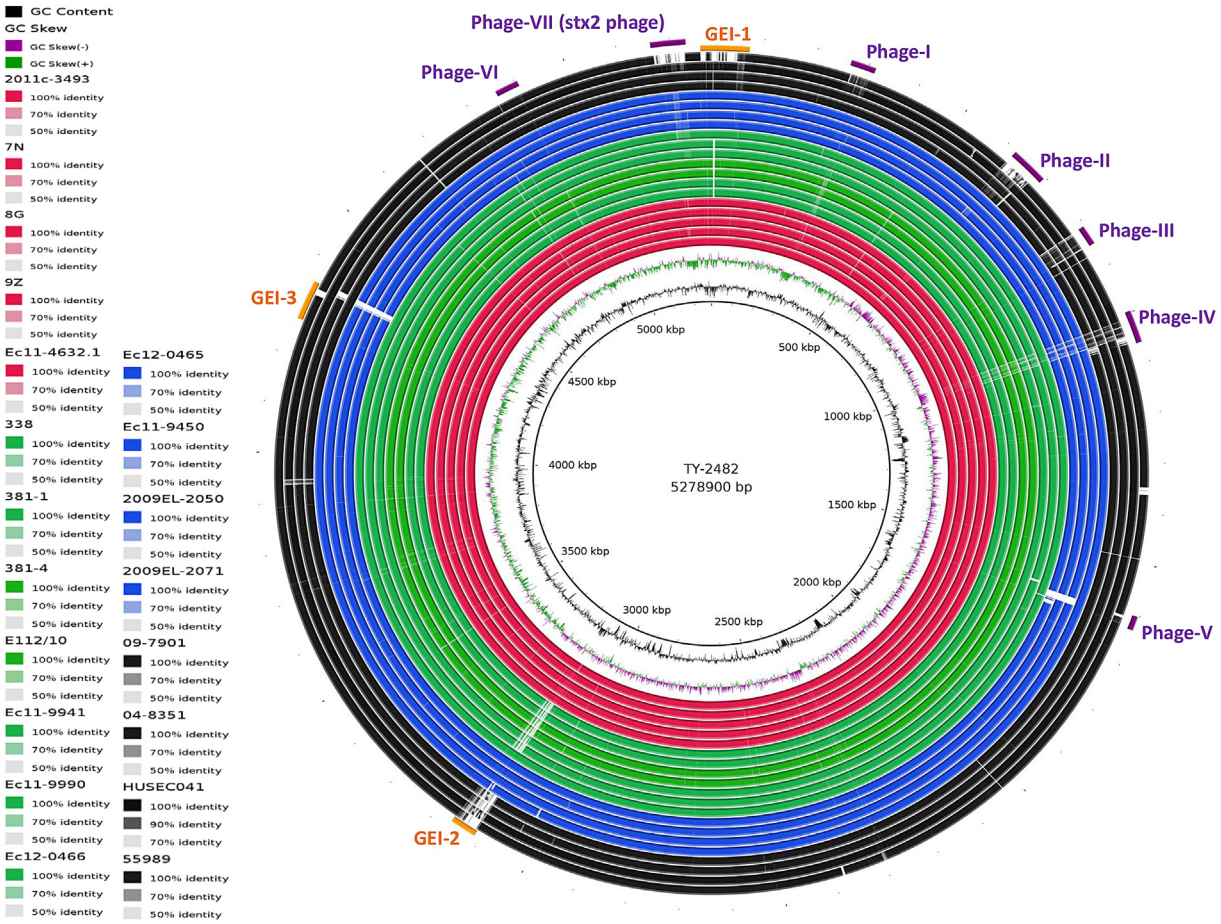


**FIG. 3.** Genomic comparison of *Escherichia coli* O104:H4. The core represents the chromosome of TY-2482 (taken as the reference genome and depicted as a black circle) and its GC content (indicated in black) and GC skew (indicated in purple/green) shown in three circles (in-outside), and the chromosomal position is numbered in a clockwise direction. Strains were grouped in different colours according to the phylogenetic results shown in Fig. 1. The order of strains followed the direction of the legend from 'GC Content' to '55989'. The gradients (dark, pale and white) of each colour represent the sequence identity (from 100% to 0%) between samples and reference defined by BLASTn. The prophages (purple) and genomic islands (orange) identified from reference TY-2482 were labelled by an arc.

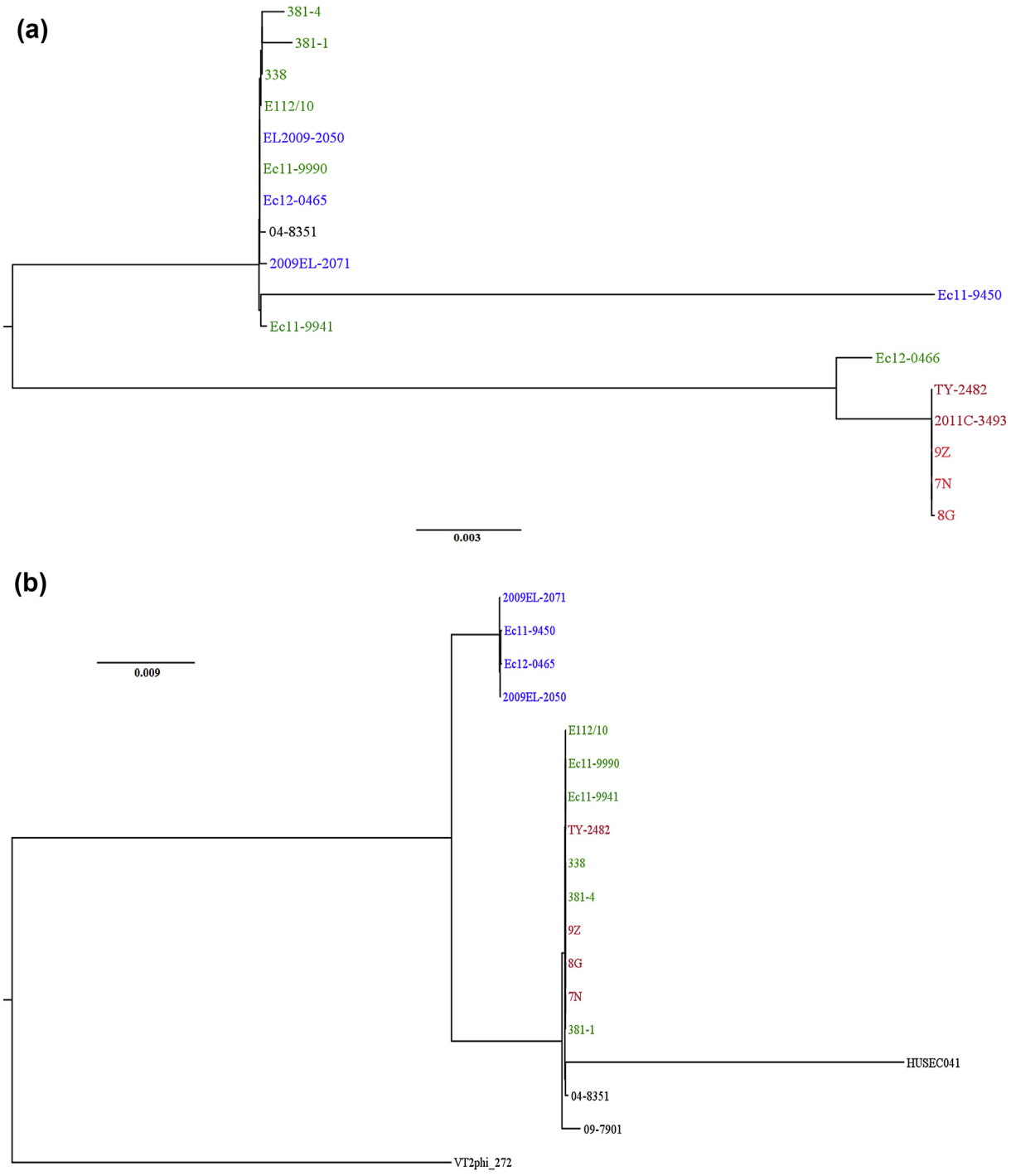


**FIG. 4.** Phylogeny of prophages uncovered from *Escherichia coli* O104:H4 STEC analysed in this study. Prophages were marked in different colours according to the phylogenetic results of their hosts shown in Fig. 1. Not all analysed strains are shown. (a) Phylogeny of phage-IV; (b) phylogeny of the Stx2-encoding prophage (phage-VII). The phage VT2phi_272 (Accession number HQ424691) was used as the outgroup.

replaced in clade B isolates by a *bor* gene (encoding a virulence factor) and another unnamed gene (encoding lytic protein), respectively (see Supporting information, Fig. S4). Additionally, Phage-V seems to be another lineage-specific prophage, which was lost by clade B isolates except 2009EL-2071 (Fig. 3). Phage-I and Phage-VI were relatively conserved among all isolates, suggesting that these two prophages were likely to be present in a common ancestor.

*Genomic islands.* Multiple GEIs were detected in each of the isolates. Here we focus on several highly diverse GEIs only (Fig. 3). One such region, referred to as GEI-1, contains the *mch* operon (microcin H47 biosynthesis), *iha* (adhesin), *ter* operon (tellurium resistance), *ag43* (antigen 43) and *yeeV-yeeU* pair (toxin–antitoxin system). This GEI was found in all isolates except in the progenitor strain 55989. However, a significant deletion with the loss of multiple genes of the *mch* operon (i.e. *mchB* and *mchC* involved in microcin biosynthesis) was found in clade A isolates with exception of Ec12-0466 (Fig. 3 and see Supporting information, Fig. S5), again consistent with its outlier position in the core-genome ML tree.

GEI-2 contains MDR genes (*dfrA7*, *sul2*, *ebr*, *strA*, *strB*, *mer operon*, *tetA*) as well as the virulence genes *ag43* and *yeeV-yeeU* pair. This region was detected in all outbreak isolates as well as clade A and B isolates, but not in any isolates of the historical clade. Syntenic analysis revealed that the structure of GEI-2 was largely conserved on the intra-clade level, but highly diverse on the inter-clade level (Fig. 3 and see Supporting information, Fig. S6) supporting the core-genome phylogeny. A large deletion including almost all of the resistance genes (*sul2*, *strA*, *strB*, *mer operon*, *tetA*) occurred in clade A isolates with the exception of Ec12-0466, again consistent with its outlier position. In contrast, all MDR genes were maintained in clade B isolates except 2009EL-2071, in which the *mer operon* and *tetA* were deleted (Fig. S6). This finding suggests a differential antibiotic selection between clade B and the outbreak clade compared with clade A (except Ec12-0466).

The third diverse GEI, named GEI-3, mainly contains the type VI secretion system (T6SS) and an incomplete prophage. A consistent deletion occurred in the region of the incomplete prophage in clade B isolates as well as in the historical strain 04-8351, whereas a different deletion within the same region occurred in 55989 (Fig. 3).

## Discussion

STEC O104:H4 has attained significant public health importance; however, little is known about the population history of the clone that caused the large outbreak in Germany in 2011. In this study, we comprehensively investigated the genomes of 23 STEC O104:H4 isolates, including previously reported outbreak- and non-outbreak-related isolates, in more detail to elucidate their evolutionary past. In accordance with our findings we propose a model as illustrated in Fig. 5. This allows a more detailed understanding of the steps that have led to the emergence of the STEC O104:H4 outbreak clone [15,18,19]. Our model reveals that the STEC O104:H4 population diversified into multiple lineages, of which two (clade A and B) derived from a recent common ancestor shared by the outbreak clone. This is the first time that two additional clones that have so far not been associated with any outbreak have been shown to share close evolutionary relationship with the 2011 outbreak clone. We presume that the three clades may represent the most successful descendants of STEC O104:H4 to date because of their present abundance among ascertained clinical cases. Notably, we identified eight canonical SNPs within coding regions that are able to unambiguously distinguish all of the 2011 outbreak isolates from the remaining population. This finding may help to set up clinical diagnostics tools (i.e. real-time PCR) to support early identification and appropriate infection control and public health measures. Additionally, seven of the eight SNPs are non-synonymous, mostly located within genes whose products were involved in metabolisms (i.e. ferredoxin-type protein NapG, uracil phosphoribosyltransferase protein and rhamnulokinase). It would be worth investigating the role of these SNPs with respect to positive selection of the outbreak clone.

Our model also describes a set of lineage-specific epidemiological markers of STEC O104:H4, some of them show the hallmarks of genomic adaptation. These findings may be helpful to identify the driving forces that lead to the diversification of STEC O104:H4. One of the obvious diversities is antibiotics, and this is supported by two observations. First, an ESBL-producing ($bla_{CTX\text{-}M\text{-}15}$) plasmid pTY1 was exclusively detected in the 2011 outbreak isolates and appeared to be relatively stable, i.e. there are no reports of pTY1 loss yet to our knowledge. This is consistent with previous studies [15,18]. Second, GEI-2 was only found in the three major clades (outbreak, A and B), and the region containing MDR genes within GEI-2 was lost in the clade A isolates (except isolate Ec12-0466). Both findings suggest that diverse antibiotic selective pressures may have shaped the evolution of STEC O104:H4. However, one may argue the influence of antibiotic-driven evolution of STEC as conventional guidelines discourage the use of antibiotics in the management of clinical STEC infections [20], but it cannot be ignored that many patients with diarrhoea receive empirical antibiotic therapy from their physicians [4]. Another driving force can be related to the niche competition. Isolates of clade A lost multiple microcin-biosynthesis genes within GEI-1. Microcin is a bactericidal antibiotic involved in competitive exclusion of other bacteria to form nutritionally restricted niches. Therefore, any habitat switch would have impact on the divergence between clade A and the other clades.

Although some other lineage-specific MGEs indicated in our model cannot directly be related to any ecological constraints, valuable information can still be extracted from our findings. For instance, it is unclear whether the replacement of pAA

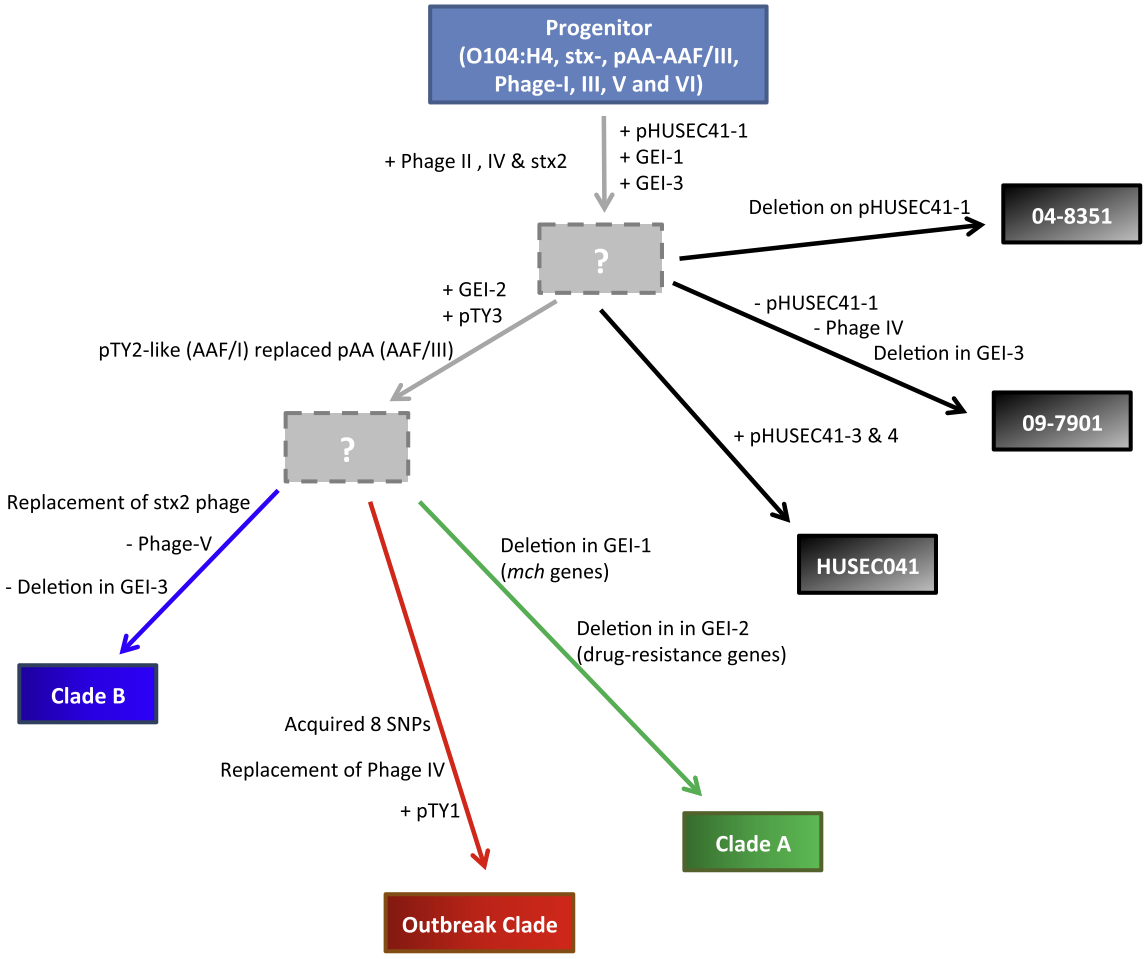


**FIG. 5.** The evolutionary model for STEC O104:H4. Populations are grouped in different colours according to the phylogenetic results shown in Fig. 1. The grey boxes with dashed outline represent hypothetical populations not identified yet. The symbol '+' and '–' represents gain and loss of mobile genetic elements, respectively. Not all events of mobile genetic element changes observed in this study are shown here.

(AAF/III) by pTY2 (AAF/I) occurring in the three major clades confers any fitness; however, our observations indicate that pTY2 may not be crucial in adhesion/colonization of the outbreak clone, i.e. the outbreak isolate 9Z lost the global regulator AggR. In fact, frequent loss of pTY2 during infection progression was detected previously from other outbreak isolates [21,22]. This differs markedly from the pAA of prototypical EAEC, which is so stable that a 1-kb fragment of this plasmid has been widely used as a sensitive and specific diagnostic marker [23]. Notably, a more recent study suggested that the pTY2 plasmid may be dispensable for the adhesion/colonization ability of STEC O104:H4 *in vivo*. This is based on observations that the overall abundance and intestinal distribution of the plasmid-devoid strains were indistinguishable from the wild-type strain in a rabbit infection model [24]. However, we cannot exclude that in humans the plasmid may be lost in the course of the infection but may still be crucial during the early phase of an infection. The cryptic plasmid pTY3 is one of the smallest plasmids found in *E. coli*, containing only a gene *repA* encoding the plasmid replication protein. Our model suggests that the hypothetical progenitor of the three major clades acquired the small cryptic plasmid after splitting from the historical clade (Fig. 5). The origin of the small cryptic plasmid is difficult to track because of its broad host range (Table S1; [25]). A recent study suggests that a pTY3-like plasmid pSERB2 (GenBank accession number: NG_036178) frequently co-transforms with an IncI1 pHUSEC-1-like plasmid (GenBank accession number: NG_035985) carrying a type IV pilus system [26]. Both plasmids have been associated with an atypical EAEC strain and are necessary for adherence to abiotic surfaces, which is required for fully mature biofilm formation in those strains [26]. We therefore speculate that pTY3 might co-transform with pTY2 acquired by STEC O104:H4, which may contribute to the pathogenicity of STEC O104:H4.

It is unclear what caused the replacement of Stx2-encoding prophages in clade B, which has also previously been reported for the two Georgia isolates 2009EL-2050 and 2009EL-2071 [27]. However, a recent study demonstrated experimentally that Stx2-encoding prophages from the 2011 German outbreak strains are completely identical to that of HUSEC041, but distinct from those from the two Georgia isolates with respect to host range and superinfection susceptibility [28]. Beutin et al. found that the replaced Stx2-encoding phage can only infect the Georgia isolates but not others [28]. Together with the core-genome (Fig. 1) and Stx2-encoding phage (Fig. 4b) phylogeny shown in this study, we suspect that a replacement event of the Stx2-encoding prophages would have occurred within clade B after the lineage split. Whether the other lineage-specific events, like the replacement of Phage-IV and acquirement of pTY3 in outbreak clone as well as the loss of Phage-V by clade B, were caused by additional ecological forces remains to be resolved.

When accepting our model one should be aware of the fact that the genome of STEC O104:H4 is rather dynamic. Moreover, certain isolates may be able to evolve much more rapidly than others within the same clade. For example, the accessory genome (i.e. GEI-1, GEI-2 and Phage-IV) of the clade A isolate Ec12-0466 was more closely related to the outbreak isolates whereas its core genome is more closely related to other clade A isolates. Rapid gain or loss of plasmids occurred in the three 2013 Dutch isolates even though they are clonal. Additionally, epidemiological data suggest that isolates of clade A and B are circulating in different regions, which may result from overlooked transmission events between these regions, i.e. by travelling and trading.

Our investigation revealed that the genome of STEC O104:H4 is mosaic in nature, mainly as the result of the frequent loss or gain of MGEs on very short evolutionary time scales. We also suggest multiple ecological constraints that may have shaped the phylogeny of STEC O104:H4. Our findings further support the hypothesis that STEC O104:H4 might have evolved to public health importance from EAEC by exploiting a rather specific cocktail of MGEs [29]. This highlights the possibility that further outbreaks could be triggered if strains attain novel combinations of MGEs. Therefore, molecular surveillance on STEC O104:H4 is necessary for early identification of the putative outbreak strains, especially in regions where they are frequently recovered from patients.

## Transparency declaration

The authors declare that they have no conflicts of interest.

## Acknowledgements

Barbara Kesztyüs and Wim Niessen are thanked for their helps to obtain the Groningen isolates and Erwin Raangs for help in obtaining the next-generation sequencing data.

## Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.cmi.2014.12.009.

## References

[1] Frank C, Werber D, Cramer JP, Askar M, Faber M, an der Heiden M, et al. Epidemic profile of Shiga-toxin-producing *Escherichia coli* O104:H4 outbreak in Germany. N Engl J Med 2011;365:1771–80.
[2] Rohde H, Qin J, Cui Y, Li D, Loman NJ, Hentschke M, et al. Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. N Engl J Med 2011;365:718–24.
[3] Bielaszewska M, Mellmann A, Zhang W, Köck R, Fruth A, Bauwens A, et al. Characterisation of the *Escherichia coli* strain associated with an outbreak of haemolytic uraemic syndrome in Germany, 2011: a microbiological study. Lancet Infect Dis 2011;11:671–7.
[4] Ishii Y, Kimura S, Alba J, Shiroto K, Otsuka M, Hashizume N, et al. Extended-spectrum beta-lactamase-producing Shiga toxin gene (Stx1)-positive *Escherichia coli* O26:H11: a new concern. J Clin Microbiol 2005;43:1072–5.
[5] Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, et al. Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. N Engl J Med 2011;365:709–17.
[6] Brzuszkiewicz E, Thürmer A, Schuldes J, Leimbach A, Liesegang H, Meyer FD, et al. Genome sequence analyses of two isolates from the recent *Escherichia coli* outbreak in Germany reveal the emergence of a new pathotype: Entero-Aggregative-Haemorrhagic *Escherichia coli* (EAHEC). Arch Microbiol 2011;193:883–91.
[7] Grad YH, Lipsitch M, Feldgarden M, Arachchi HM, Cerqueira GC, Fitzgerald M, et al. Genomic epidemiology of the *Escherichia coli* O104:H4 outbreaks in Europe 2012. Proc Natl Acad Sci U S A 2011;109:3065–70.
[8] Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics 2009;25:1968–9.
[9] Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST server: rapid annotations using subsystems technology. BMC Genomics 2008;9:75.
[10] Darling AE, Mau B, Perna NT. ProgressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 2010;5:e11147.
[11] Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. PHAST: a fast phage search tool. Nucleic Acids Res 2011;39:W347–52.
[12] Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics 2011;12:402.
[13] Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, et al. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet 2009;5:e1000344.
[14] Kunne C, Billion A, Mshana SE, Schmiedel J, Domann E, Hossain H, et al. Complete sequences of plasmids from the hemolytic-uremic

syndrome-associated *Escherichia coli* strain HUSEC41. J Bacteriol 2012;194:532–3.

[15] Ahmed SA, Awosika J, Baldwin C, Bishop-Lilly KA, Biswas B, Broomall S, et al. Genomic comparison of *Escherichia coli* O104:H4 isolates from 2009 and 2011 reveals plasmid, and prophage heterogeneity, including shiga toxin encoding phage stx2. PLoS One 2012;7:e48228.

[16] Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 2006;22:2688–90.

[17] Guy L, Jernberg C, Ivarsson S, Hedenström I, Engstrand L, Andersson SG. Genomic diversity of the 2011 European outbreaks of *Escherichia coli* O104:H4. Proc Natl Acad Sci U S A 2012;109:E3627–8.

[18] Grad YH, Godfrey P, Cerquiera GC, Mariani-Kurkdjian P, Gouali M, Bingen E, et al. Comparative genomics of recent Shiga toxin-producing *Escherichia coli* O104:H4: short-term evolution of an emerging pathogen. MBio 2013;4:e00452–12.

[19] Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, et al. Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. PLoS One 2011;6:e22751.

[20] Paton JC, Paton AW. Pathogenesis and diagnosis of Shiga toxin-producing *Escherichia coli* infections. Clin Microbiol Rev 1998;11:450–79.

[21] Zangari T, Melton-Celsa AR, Panda A, Boisen N, Smith MA, Tatarov I, et al. Virulence of the Shiga toxin type 2-expressing *Escherichia coli* O104:H4 German outbreak isolate in two animal models. Infect Immun 2013;81:1562–74.

[22] Zhang W, Bielaszewska M, Kunsmann L, Mellmann A, Bauwens A, Köck R, et al. Lability of the pAA virulence plasmid in *Escherichia coli* O104:H4: implications for virulence in humans. PLoS One 2013;8:e66717.

[23] Baudry B, Savarino SJ, Vial P, Kaper JB, Levine MM. A sensitive and specific DNA probe to identify enteroaggregative *Escherichia coli*, a recently discovered diarrheal pathogen. J Infect Dis 1990;161:1249–51.

[24] Munera D, Ritchie JM, Hatzios SK, Bronson R, Fang G, Schadt EE, et al. Autotransporters but not pAA are critical for rabbit colonization by Shiga toxin-producing *Escherichia coli* O104:H4. Nat Commun 2014;5:3080.

[25] Srivastava P, Nath N, Deb JK. Characterization of broad host range cryptic plasmid pCR1 from *Corynebacterium renale*. Plasmid 2006;56:24–34.

[26] Dudley EG, Abe C, Ghigo JM, Latour-Lambert P, Hormazabal JC, Nataro JP. An IncI1 plasmid contributes to the adherence of the atypical enteroaggregative *Escherichia coli* strain C1096 to cultured cells and abiotic surfaces. Infect Immun 2006;74:2102–14.

[27] Guy L, Jernberg C, Arvén Norling J, Ivarsson S, Hedenström I, Melefors Ö, et al. Adaptive mutations and replacements of virulence traits in the *Escherichia coli* O104:H4 outbreak population. PLoS One 2013;8:e63027.

[28] Beutin L, Hammerl JA, Strauch E, Reetz J, Dieckmann R, Kelner-Burgos Y, et al. Spread of a distinct Stx2-encoding phage prototype among *Escherichia coli* O104:H4 strains from outbreaks in Germany, Norway, and Georgia. J Virol 2012;86:10444–55.

[29] Baquero F, Tobes R. Bloody coli: a gene cocktail in *Escherichia coli* O104:H4. MBio 2013;4:e00066–13.

[30] Jourdan-da Silva N, Watrin M, Weill FX, King LA, Gouali M, Mailles A, et al. Outbreak of haemolytic uraemic syndrome due to Shiga toxin-producing *Escherichia coli* O104:H4 among French tourists returning from Turkey, September 2011. Euro Surveill 2012;17. pii: 20065.

[31] Scheutz F, Nielsen EM, Frimodt-Møller J, Boisen N, Morabito S, Tozzoli R, et al. Characteristics of the enteroaggregative Shiga toxin/verotoxin-producing *Escherichia coli* O104:H4 strain causing the outbreak of haemolytic uraemic syndrome in Germany, May to June 2011. Euro Surveill 2011;16. pii: 19889.

[32] Monecke S, Mariani-Kurkdjian P, Bingen E, Weill FX, Balière C, Slickers P, et al. Presence of enterohemorrhagic *Escherichia coli* ST678/O104:H4 in France prior to 2011. Appl Environ Microbiol 2011;77:8784–6.

[33] Mellmann A, Bielaszewska M, Köck R, Friedrich AW, Fruth A, Middendorf B, et al. Analysis of collection of hemolytic uremic syndrome-associated enterohemorrhagic *Escherichia coli*. Emerg Infect Dis 2008;14:1287–90.